Comp 598: Assignment 2

Protein Structure and System Biology

Due on April 13th, 2014.

- To some extent, collaborations are allowed, but you must indicate the name of all collaborators (including instructors) on your answers. Uncredited collaborations will be penalized.
- Unless specified, all answers must be justified.
- Partial answers will receive credits.
- Answers should be submitted electronically to the instructor.

Exercise 1 (10 points) Explain why secondary structure prediction methods are more accurate at predicting α-helices than β-sheets (N.B.: Provide a detailed answer).

Exercise 2 (35 points) We aim to develop a simple method to predict protein secondary structures. We will restrict the scope of this work to the prediction of residues in α-helices (noted `H`) or coil regions (noted `C`). Several propensity scale have been proposed for α-helices. Here, we will use the scale proposed by C.N. Pace and J.M. Scholtz in `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1299714/`. We will benchmark our techniques on the myoglobin: `http://www.rcsb.org/pdb/explore/explore.do?structureId=1mbn` (extract the primary and secondary structure from the Protein Data Bank record).

1. Implement an algorithm that assigns a secondary structure type to each residue of an input sequence from the propensity scale introduced above. The program will require users to input a protein sequence ω and a threshold value λ. Residues with a propensity value lower than λ will be predicted to belong to an α-helix secondary structure.

2. Implement a program that compare a "real" secondary structure with a predicted one, and calculate the true positive rate (TPR) and false positive rate (FPR), respectively defined as $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$ where TP are True Positive, FN are False Negatives, FP are False Positives, and TN are True Negatives.

3. Use these programs to plot the receiver operating characteristic (ROC) curve and calculate the area under the curve (AUC) (`http://en.wikipedia.org/wiki/Receiver_operating_characteristic`). Hint: You will vary the threshold λ to determine the coordinate of the points delimitating the hull of the ROC curve.

4. Improve your $alpha$-helix predictor. You will incorporate in your algorithm, a signal from sequence neighbour residues. In particular, the propensity value associate to a residue will now be the average of the propensity values of all residues located 4 positions before or after the current index.

5. Repeat the procedure of the third item, compare and discuss the performance of the two versions of the predictor.

Exercise 3 (25 points) We want to predict β-sheets from a residue contact matrix. We provide a contact map from protein GB1 at `http://www.cs.mcgill.ca/~jeromew/comp598/data/2QMT.ct`.

1. Propose and implement an algorithm that detect all parallel and anti-parallel β-strand pairs from a contact map. Here, a β-strand pair must have at least 4 consecutive residue contact. Your program will output a list of β-strand pairs using the following format: Orientation (i.e. `A` for anti-parallel or `P` for parallel), Length (i.e. number of contacts), indices of the first contact (i.e. contact with the lowest sequence index).

2. Write the pseudo-code of a greedy algorithm that selects the longest β-strand pair and then extend it with β-strand pairs that are compatible with those previously selected. New β-strand pairs must have one β-strand that overlaps by at least 4 residues with a β-strand in the previous β-sheet. The other β-strand must not intersect with any other β-strand. Each β-strand can pair at most twice. β-strand pairs with the largest number of contacts must be inserted first.

3. Implement the algorithm and apply it on the contact map of protein GB1.

Exercise 4 (30 points) We will analyze the result of our MD simulation. We simulated a system for 2000 pico seconds with a mutated version of amylin. Your job is to create the RMSD graph for this simulation along with a short movie showing the protein in action.

To do this, you'll have to:

1. Download the 'npt-nopr.tpr' and the 'npt-nopr.trr' files from the dropbox link that was sent to you. These two files store the result of the simulation trajectory of every atom.

2. Use the 2 commands on slide 8 of the tutorial powerpoint that was given to you (found on http://cs.mcgill.ca/ ms-maou/MD). The first command will create the RMSD graph, and the second will create the movie.

3. Generate a graph from a software that opens .xvg extensions

4. Open your molecule-movie.pdb file in PYMOL and click "File->save as->Movie->choose .avi or .mov